

Binary, Shortened Projective Reed Muller Codes for Coded Private Information Retrieval

Myna Vajha, Vinayak Ramkumar, and P. Vijay Kumar

Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore.

Email: {myna, vinram, vijay}@ece.iisc.ernet.in

Abstract

The notion of a Private Information Retrieval (PIR) code was recently introduced by Fazeli, Vardy and Yaakobi [1] who showed that this class of codes permit PIR at reduced levels of storage overhead in comparison with replicated-server PIR. In the present paper, the construction of an (n, k) τ -server binary, linear PIR code having parameters $n = \sum_{i=0}^{\ell} \binom{m}{i}$, $k = \binom{m}{\ell}$ and $\tau = 2^\ell$ is presented. These codes are obtained through homogeneous-polynomial evaluation and correspond to the binary, Projective Reed Muller (PRM) code. The construction can be extended to yield PIR codes for any τ of the form 2^ℓ , $2^\ell - 1$ and any value of k , through a combination of single-symbol puncturing and shortening of the PRM code. Each of these code constructions above, have smaller storage overhead in comparison with other PIR codes appearing in the literature.

For the particular case of $\tau = 3, 4$, we show that the codes constructed here are optimal, systematic PIR codes by providing an improved lower bound on the block length $n(k, \tau)$ of a systematic PIR code. It follows from a result by Vardy and Yaakobi [2], that these codes also yield optimal, systematic primitive multi-set $(n, k, \tau)_B$ batch codes for $\tau = 3, 4$. The PIR code constructions presented here also yield upper bounds on the generalized Hamming weights of binary PRM codes.

Index Terms

PIR codes, private information retrieval, replicated-server PIR, Projective Reed-Muller code, shortened code.

I. INTRODUCTION

Private Information Retrieval (PIR) refers to the retrieval of data from a database without revealing information about the data being retrieved to the servers. Considering Q_J as the set of queries sent to the database in order to retrieve a symbol X_J whose index in the database is given by random variable J , we require the mutual information $I(Q_J; J)$ to be zero. The PIR problem was first introduced by Chor et al. in [3] who showed that the communication complexity needs be of order $\Omega(B)$ when a single server with database of size B is employed. To reduce communication complexity, the authors of [3] introduced the model of non-communicating servers that store replicas of the same database and proposed algorithms for achieving PIR. On restricting to replicated server setting, the PIR algorithms require storage overhead to be ≥ 2 . In [4] the idea of erasure coding across PIR servers was introduced, but the metric of interest there was the amount of data downloaded and not the storage overhead. In [5], PIR schemes based on locally-decodable codes are discussed. Coded-PIR was further explored in [6] in which the trade-off between download and storage overhead is studied.

In [1], [7] Fazeli, Vardy and Yaakobi came up with the notion of PIR codes to achieve low storage overhead. Given an (n, k) τ -server PIR code, where n denotes the number of servers with each server storing $\frac{B}{k}$ coded symbols, the authors provide an algorithm to achieve PIR using any existing τ -replicated server protocol. An (n, k) τ -server PIR code, is an (n, k) linear code such that for every message symbol m_i , $i \in [k]$, there are τ disjoint recovery sets $R_{it} \forall t \in [\tau]$ such that: $m_i = \sum_{j \in R_{it}} c_j \forall t \in [\tau]$, where $\underline{c} = (c_1, \dots, c_n)$ is a codeword. By disjoint recovery sets, it is meant that $R_{it_1} \cap R_{it_2} = \emptyset$ whenever $t_1 \neq t_2$ and for any $i \in [k]$. For a PIR code with $c_j = m_i$, the singleton set $\{j\}$ can itself act as a recovery set for m_i . Thus in the case of a systematic PIR code, every message symbol has at least one recovery set of size 1.

Myna would like to thank the support of Visvesvaraya PhD Scheme for Electronics & IT awarded by DEITY, Govt. of India. P. V. Kumar is also an Adjunct Research Professor at the University of Southern California. His research is supported in part by the National Science Foundation under Grant No. 1421848 and in part by the joint UGC-ISF research program.

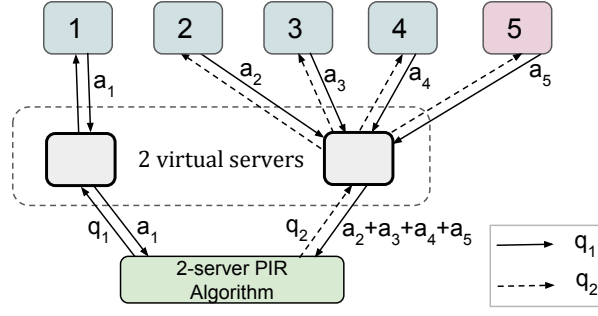


Fig. 1. The working of the $(5, 4)$ 2-server PIR code described in Example 1 is illustrated here.

Example 1: The working of a PIR code (see [1]) is explained through an example that is illustrated in Fig. 1. A database of size $4B$ symbols is partitioned into the 4 subsets $\{x_{ij} \mid j \in [B]\}_{i=1}^4$ and the i th subset is stored on the server numbered i . The 5th server stores B symbols, each of which is the modulo-2 sum of the corresponding contents of the 4 servers. Let Q and A be the query and answer functions for a replicated 2-server PIR algorithm. In order to retrieve x_{1j} , queries $q_1 = Q(1, j)$, $q_2 = Q(2, j)$ are generated. The queries q_1, q_2 are respectively sent to the server sets corresponding to the two recovery sets $R_{11} = \{1\}$ and $R_{12} = \{2, 3, 4, 5\}$ for message symbol 1 in the PIR code. Let $\{a_i\}_{i=1}^5$ be the corresponding responses, where $a_1 = A(q_1, x_1)$ and $a_i = A(q_2, x_i), \forall i \in R_{12}$. This algorithm assumes linearity of function A in its second parameter, that results in $\sum_{i \in R_{12}} A(q_2, x_i) = A(q_2, x_2 + x_3 + x_4 + x_5) = A(q_2, x_1)$. The PIR algorithm determines x_{1j} from $A(q_1, x_1) = a_1$ and $A(q_2, x_1)$.

In [1], several PIR code constructions were proposed and connections with locally recoverable codes were made. In [8], the authors prove a $\Omega(\sqrt{k})$ lower bound on the redundancy of an (n, k) τ -server PIR code and showed that this matches with the $\mathcal{O}(\sqrt{k})$ upper bound that follows from the PIR constructions in [1]. PIR array codes are also introduced in [1], and [9], [10] are two recent works in that direction. In [2], primitive multi-set batch code constructions were given using PIR codes. A (n, k) linear code is called a $(n, k, \tau)_B$ primitive multi-set batch code if for any collection of τ message symbols $\underline{i} = (i_1, \dots, i_\tau)$ with repetition permitted, for all $t \in [\tau]$, there exists a recovery set R_t for symbol i_t , such that $R_{t_1} \cap R_{t_2} = \emptyset$. In [2] it is shown that for $\tau = 3, 4$ optimal, systematic, PIR codes are also optimal, systematic primitive multi-set batch codes.

A. Contributions

In the present paper, constructions for systematic PIR codes for τ of the form $2^\ell, 2^\ell - 1$, are provided by appropriately shortening a PRM code and it is shown that these codes have lower storage overhead (smaller block lengths) in comparison with known codes[1]. A lower bound on the block length of a systematic PIR code is presented and for $\tau = 3, 4$, the codes constructed here, are shown to be optimal with respect to this bound.

B. Organization

Section II presents a primer on Reed Muller (RM) codes. Binary PRM codes are introduced in Section III and it is shown that this class yields efficient PIR codes. In Section IV, a support set viewpoint of PRM codes is presented and used in Section V, to provide constructions of PIR codes for any k . Upper bounds on the generalized Hamming weights of binary PRM codes, obtained as a by-product, appear in Section V. In Section VI, an improved lower bound for systematic PIR codes is presented and used in Section VII, to prove optimality of the constructions for $\tau = 3, 4$.

We use $[a, b]$ to denote $\{a, a+1, \dots, b-1, b\}$, $[a] = [1, a]$, $(a, b) = [a, b] \setminus \{a\}$ and $[a, b) = [a, b] \setminus \{b\}$.

II. REED MULLER CODE

A codeword in a RM code $RM(r, m)$ [11] is a vector of 2^m evaluations of a polynomial

$$f(x_1, \dots, x_m) = \sum_{S \subseteq [m], |S| \leq r} a_S \prod_{i \in S} x_i, \quad a_S \in \mathbb{F}_2, \quad (1)$$

of degree $\leq r$ over \mathbb{F}_2 , in the m binary variables x_i . The coefficients a_S represent the information symbols. The $RM(r, m)$ code has parameters: $n = 2^m$ and $k = \sum_{i=0}^r \binom{m}{i}$.

A sequential decoding algorithm to recover message symbols is provided in [12]. The coefficients corresponding to the highest-degree monomials are decoded first according to:

$$a_R = \sum_{x_R \in \mathbb{F}_2^r} f(x_R, \underline{b}) \text{ for any } R \subseteq [m] \text{ and } |R| = r, \quad (2)$$

where $x_R \in \mathbb{F}_2^r$, refers to the collection of variables $(x_i | \forall i \in R)$ and $\underline{b} \in \mathbb{F}_2^{m-r}$ corresponds to a particular value of $x_{[m] \setminus R}$. There are 2^{m-r} possible values \underline{b} can take resulting in 2^{m-r} recovery equations. On considering recovery equations corresponding to \underline{b}_1 and \underline{b}_2 where $\underline{b}_1 \neq \underline{b}_2$ for a given message symbol a_R , it can be seen that the indices of code symbols involved are disjoint. Therefore any a_R for all $R \subseteq [m]$ and $|R| = r$, can be recovered from 2^{m-r} disjoint recovery sets. Having recovered the coefficient of the highest-degree monomial terms, the contribution of these highest-degree terms is then subtracted out, leaving us with a Boolean function of lesser degree and the process is then repeated with this lesser degree.

III. THE PROJECTIVE REED MULLER CODE CONSTRUCTION

On account of the sequential nature of the recovery algorithm, more information is needed during the recovery of lower-degree coefficients in comparison with the coefficients of the degree- r terms. To gain access to a message symbol corresponding to a degree $i < r$ term, all the message symbols corresponding to degree $> i$ have to be previously determined.

Clearly, this can be avoided if the polynomials appearing in (1), were restricted to be homogeneous, i.e., the coefficients of all the lower-degree monomial terms are set equal to zero. The restriction of evaluation to homogeneous polynomials takes us from the setting of conventional and affine RM codes to the setting of Projective Reed-Muller (PRM) codes.

Projective Reed-Muller (PRM) codes over the field \mathbb{F}_q were introduced in [13]. A codeword in the $PRM(r, m-1)$ code corresponds to evaluations of a homogeneous polynomial of degree r at a specifically-chosen representative of each of the points in the projective space $\mathbb{P}^{m-1}(\mathbb{F}_q)$. We note however, that in the projective space $\mathbb{P}^{m-1}(\mathbb{F}_2)$, each point in projective space has just a single unique representative with m components. While the block length of a binary $PRM(r, m-1)$ code is nominally equal to $2^m - 1$, the evaluation of a homogeneous polynomial of degree r at any coordinate \underline{x} with $\text{supp}(\underline{x}) < r$ gives the value 0. Hence, these coordinates can be deleted from the binary $PRM(r, m-1)$ code to obtain a shortened version. From now on when we refer to $PRM(r, m-1)$ code, its the shortened binary version that we refer to. It follows that the code $PRM(r, m-1)$ has block length $n = \sum_{i=r}^m \binom{m}{i}$ and dimension $k = \binom{m}{r}$.

Each message symbol in the PRM code can be recovered by the same method used to recover degree- r terms in the Reed Muller code as shown in (2). In the recovery equation for message symbol a_R given by the vector \underline{b} , it can be verified that there is at least one element in the summation in (2). This ensures that there are $\tau = 2^{m-r}$ disjoint recovery sets for the retrieval of any message symbol. Hence the $PRM(r, m-1)$ code is a (n, k) , τ -server PIR code, where

$$n = \sum_{i=r}^m \binom{m}{i}, \quad k = \binom{m}{r} \text{ and } \tau = 2^{m-r}.$$

Additionally, the recovery equation corresponding to $\underline{b} = \underline{0}$ for any message symbol a_R , gives us $a_R = f(\underline{1}_R)$, where $\underline{1}_R$ is a binary vector with support set R . This establishes that the code $PRM(r, m-1)$ is a systematic code.

Example 2: Consider the code $PRM(2, 3)$. This code has parameters $(n = 11, k = 6, \tau = 4)$. A code vector in $PRM(2, 3)$ corresponds to the evaluation of polynomials of form $f(\underline{x}) = a_{12}x_1x_2 + a_{13}x_1x_3 + a_{14}x_1x_4 + a_{23}x_2x_3 + a_{24}x_2x_4 + a_{34}x_3x_4$ of degree 2 in 4 variables at points $\underline{x} = (x_1, x_2, x_3, x_4)$ such that $w_H(\underline{x}) \geq 2$. Next, consider the recovery of the coefficient a_{12} , of x_1x_2 . This coefficient can be recovered by fixing (b_3, b_4) and summing over

the support of the corresponding recovery sets as shown below. There are 4 possible values of (b_3, b_4) and hence 4 disjoint recovery sets for a_{12} .

$$\begin{aligned}
a_{12} &= \sum_{x_1, x_2} f(x_1 x_2 b_3 b_4) \\
&= f(1100) \\
&= f(0110) + f(1010) + f(1110) \\
&= f(0101) + f(1001) + f(1101) \\
&= f(0011) + f(0111) + f(1011) + f(1111).
\end{aligned}$$

Generator matrix (permuted) for the $PRM(2, 3)$ code is given by:

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

The PRM codes have in general, non-uniform information-symbol locality. For instance, in the example above, there are 2 sets with locality 3 and 1 set with locality 4. Each recovery set $R_{\underline{b}}$ is naturally associated to a specific vector $\underline{b} \in \mathbb{F}_2^{m-r}$. Let w_b denote the Hamming weight of the vector \underline{b} . There are $\binom{m-r}{w_b}$ recovery sets with cardinality $R_{\underline{b}}$ and

$$|R_{\underline{b}}| = \begin{cases} \sum_{i=0}^{w_b} \binom{r}{r-w_b+i} & w_b < r, \\ 2^r & w_b \geq r. \end{cases}$$

Since $|\cup_{\underline{b} \in \mathbb{F}_2^{m-r}} R_{\underline{b}}| = n$, it follows that all code symbols participate in the τ recovery equations corresponding to each of the message symbols.

IV. SUPPORT-SET VIEWPOINT OF PRM CODES

Each code symbol of an $PRM(r, m-1)$ code, is indexed by a vector $\underline{x} \in \mathbb{F}_2^m$ with $w_H(\underline{x}) \geq r$. Since each of these vectors is uniquely represented by its support set, each code symbol can equivalently, be indexed by a subset of $[m]$ of size $\geq r$. Our aim in the next section, is to construct PIR codes for other values of k . Our approach is to consider shortened versions of the PRM code, obtained by judiciously setting certain of the message symbols to zero. When we set a certain message symbol to equal zero, the corresponding code symbol (since the code is systematic) is automatically set equal to zero. But if a set of message coefficients is set equal to zero, it turns out that certain other code symbols are forced to be equal to zero as well. This results in a shortened code having smaller block length. The shortened codes are also PIR codes for exactly the same reason as is the parent PRM code. The shorter block length makes these codes more efficient as can be seen from the table VII of the parameters of the PIR codes so constructed. We explain this last point in greater detail below.

For S a subset of $[m]$, we will for the sake of brevity, write $f(S)$ in place of $f(\underline{1}_S)$. For example, when $m = 5$, we will write $f(\{1, 2, 5\})$ in place of $f(11001)$. Next, let $R_i, \forall i \in \binom{[m]}{r}$ represent the $\binom{[m]}{r}$, r -element subsets of $[m]$. We note that for any subset $S \subseteq [m]$, we have that

$$f(S) = \sum_{\forall R_i \subseteq S} f(R_i).$$

For example, $PRM(2, 3)$ code has $f(\{1, 2, 4\}) = f(\{1, 2\}) + f(\{1, 4\}) + f(\{2, 4\})$.

It follows that if we set $f(R_i) = 0$, by setting the corresponding message coefficients to be equal to zero, $\forall R_i \subseteq S$, then $f(S) = 0$. Thus if we shorten the PRM code by setting all message coefficients corresponding to r -element subsets of a fixed set S to zero, then the shortening process will result in the deletion of the coordinate corresponding to the support set S as well.

V. CONSTRUCTIONS FOR ANY k AND $\tau = 2^\ell, 2^\ell - 1$

In this section we provide constructions for τ of the form 2^ℓ for any k . Each of these codes will also turn out to be systematic. It is straightforward to show (see [1]) that if a systematic (n, k) , τ -server PIR code is punctured by deleting a parity-check symbol, one will obtain a systematic $(n - 1, k)$, $(\tau - 1)$ -server PIR code. Thus our constructions for (n, k) , 2^ℓ -server PIR codes, can be punctured to yield constructions for $\tau = 2^\ell - 1$ as well.

In this section, we will show how one can make use of the support-set viewpoint of a PRM code to shorten the code to obtain PIR codes for values of k other than of the form $\binom{m}{\ell}$. To construct a PIR code for $k \in ((\binom{m-1}{\ell}, \binom{m}{\ell}))$ and $\tau = 2^\ell$, we consider a $PRM(r, m - 1)$ code, where $r = m - \ell$ and set $\gamma = \binom{m}{\ell} - k$, message symbols to zero to obtain the shortened Projective Reed Muller code $SPRM(r, m - 1, \gamma)$ for $0 \leq \gamma \leq \binom{m-1}{\ell-1}$. Considering γ' as the reduction in block length on shortening $PRM(r, m - 1)$ by γ , we get $n = \sum_{i \in [0, \ell]} \binom{m}{i} - \gamma'$. It is clear that $\gamma' \geq \gamma$.

We first show in Lemma 5.1 that irrespective of setting any of the γ message symbols to zero, τ is still retained. We then give an algorithm to judiciously pick the γ message symbols to get a block length reduction of γ' in Theorem 5.4.

Lemma 5.1: On shortening a $PRM(r, m - 1)$ code by setting any γ message symbols to zero, the resultant code retains $\tau = 2^{m-r}$ disjoint recovery sets.

Proof: Consider $f(R_j)$, $\forall j \in [\gamma]$ as the γ message symbols that are set to zero. Any recovery equation for a left out symbol $f(R_i)$ for $i \in [\gamma + 1, \binom{m}{r}]$ given below has $f(R_i \cup S)$ as an element.

$$f(R_i) = \sum_{R_0 \subseteq R_i} f(R_0 \cup S) \quad \forall S \subseteq [m] \setminus R_i.$$

It is clear to see that $f(R_i \cup S)$ cannot be deleted when $f(R_i)$ is not set to 0. This shows that for any $S \in [m] \setminus R_i$ we have at-least one element in the recovery equation, resulting in $\tau = 2^\ell$.

Theorem 5.2: For $\gamma = \binom{r+t}{r}$ for all $t \in [0, \ell - 1]$, $\gamma' = \sum_{i=0}^t \binom{r+t}{r+i}$ is possible.

Proof: Consider a $r + t$ element subset T of $[m]$ and shorten $PRM(r, m - 1)$ by setting the γ message symbols corresponding to all the r -element subsets of T as zero. By doing this, we can also delete code symbols corresponding to the subsets of T with cardinality $\geq r$. This gives a reduction of $\gamma' = \sum_{i=0}^t \binom{r+t}{r+i}$.

For the case of $PRM(2, 4)$ code, Theorem 5.2 gives the codes $SPRM(2, 4, \gamma)$ for $\gamma = 1, 3, 6$ with γ' as 1, 4, 11 respectively. On setting $t = \ell - 1$ in Theorem 5.2 we get the parameters of $SPRM(r, m - 1, \binom{m-1}{r})$ code as $k = \binom{m-1}{\ell}$, $n = \sum_{i=0}^{\ell} \binom{m-1}{i}$. These parameters are same as that of $PRM(r - 1, m - 2)$. Therefore, we do not restrict to $\gamma < \binom{m-1}{\ell}$ in the next theorems as this shortening algorithm seamlessly goes from $PRM(r, m - 1)$ to $PRM(r - 1, m - 2)$.

Theorem 5.3: For $\gamma = \sum_{i=0}^{\rho_t-1} \binom{r+t-i}{r-i}$ for any $t \in [0, \ell - 1]$ and $\rho_t \in [1, r]$, $\gamma' = \sum_{j=0}^t \sum_{i=0}^{\rho_t-1} \binom{r+t-i}{r+j-i}$ is possible.

Proof: Consider the set $S = [1, r + t + 1]$ and the $(r + t)$ -element subsets $S_i = S \setminus \{r + t + 1 - i\}$, $\forall i \in [0, r + t]$.

Consider ρ_t such $(r + t)$ -element sets $\mathbb{P} = \{S_i, \forall i \in [0, \rho_t - 1]\}$ where $\rho_t \leq r$ and shorten $PRM(r, m - 1)$ by setting message symbols corresponding to all the distinct r element subsets of sets in \mathbb{P} . This gives $\gamma = \sum_{i=0}^{\rho_t-1} \binom{r+t-i}{r-i}$.

In this case we can delete all the code symbols corresponding to subsets of sets in \mathbb{P} with cardinality $\geq r$ giving a reduction of $\gamma' = \sum_{j=0}^t \sum_{i=0}^{\rho_t-1} \binom{r+t-i}{r+j-i}$ resulting in the theorem.

For $\rho_t = 1$, Theorem 5.3 falls back to the case of Theorem 5.2. Now by picking $\rho_t = 2$ for $PRM(2, 4)$ code in Theorem 5.3 we get the $SPRM(2, 4, \gamma)$ code for $\gamma = 2, 5, 9$ with $\gamma' = 2, 7, 18$ respectively. We essentially extend the same idea in the next theorem to give constructions for any k .

Theorem 5.4: For any $\gamma \in [0, \binom{m}{\ell}]$, γ can be uniquely represented using a vector $(\rho_{\ell-1}, \dots, \rho_0)$ with $\rho_i \geq 0, \forall i \in [0, \ell-1]$ and $\sum_{i=0}^{\ell-1} \rho_i \leq r$ as

$$\gamma = \sum_{t=0}^{\ell-1} h(\rho_t, r_t, t) \quad \text{where, } h(p, r, t) = \begin{cases} \sum_{i=0}^{p-1} \binom{r+t-i}{r-i} & p > 0 \\ 0 & p = 0 \end{cases} \quad \text{and } r_t = r - \sum_{q>t}^{\ell-1} \rho_q.$$

Then for $\text{SPRM}(r, m-1, \gamma)$, reduction of

$$\gamma' = \sum_{t=0}^{\ell-1} h_1(r_t, t) \quad \text{where, } h_1(r, t) = \begin{cases} \sum_{j=0}^t \sum_{i=0}^{\rho_t-1} \binom{r+t-i}{r+j-i} & \rho_t > 0 \\ 0 & \rho_t = 0 \end{cases}$$

is possible.

Proof: Lets recursively define

$$\gamma_t = \begin{cases} \gamma & t = \ell - 1, \\ \gamma_{t+1} - h(\rho_{t+1}, r_{t+1}, t+1) & 0 \leq t < \ell - 1. \end{cases}$$

We determine ρ_t as shown below by the index $p \in [0, r_t]$ of the interval in which γ_t lies.

$$\rho_t = p \text{ such that } \gamma_t \in \left[h(p, r_t, t), h(p+1, r_t, t) \right).$$

For $t = \ell - 1$, $\gamma < h(r+1, r, \ell-1) = \binom{r+\ell}{r} = \binom{m}{r}$. One can always find an interval in which γ_t lies, otherwise we have $\gamma_t \geq h(r_t+1, r_t, t) = \binom{r_t+t+1}{r_t}$. This gives that

$$\begin{aligned} \gamma_{t+1} &\geq h(\rho_{t+1}, r_{t+1}, t+1) + \binom{r_t+t+1}{r_t} \\ &= h(\rho_{t+1}+1, r_{t+1}, t+1) \text{ \{ as } r_t = r_{t+1} - \rho_{t+1} \}. \end{aligned}$$

This is a contradiction on definition of ρ_{t+1} . So we can always find an index $p \in [0, r_t]$ for ρ_t . We start by defining the global set as $S_0^\ell = [m]$ and define $\rho_\ell = 0$. For the set S_i^j , j is the number of elements in the set. Now we recursively define sets,

$$S_i^{r+t-1} = S_{\rho_t}^{r+t} \setminus \{r_{t-1} + t - i\}, \quad \forall i \in [0, r_{t-1} + t - 1] \quad (3)$$

$\forall t \in [1, \ell]$. It is clear to see that $|S_i^j \cap S_{i'}^j| = j-1$ for the sets defined by 3. By picking ρ_t , $(r+t)$ -element sets, we get $\mathbb{P} = \{S_i^{r+t}, \forall t \in [0, \ell-1], \forall i \in [0, \rho_t-1]\}$. It can be seen that $S_i^{r+t} \not\subseteq S_{i'}^{r+t'}$ for all $t' > t$ and $i' \in [0, \rho_{t'}-1]$. Here, ρ_t corresponds to the number of $r+t$ element sets that are not already subsets of larger cardinality sets in \mathbb{P} . Now by setting all the message symbols corresponding to distinct r -element subsets of sets in \mathbb{P} to zero we get a count of γ . Now we can delete symbols corresponding to all subsets of sets in \mathbb{P} with cardinality $\geq r$. This gives us the reduction γ' as stated.

Theorem 5.3 is a special case of Theorem 5.4, where γ is represented by single weight $\underline{\rho}$ vector. This can be seen in Table I.

A. Upper bounds on generalized Hamming weights of Binary PRM codes.

The SPRM codes presented in section V also give upper bound on the generalized Hamming weights of PRM codes defined as $d_i = \min |\text{supp}(D)|$ where D is a i -dimensional sub code and $\text{supp}(D)$ is the union of support of all the vectors in D . For a binary PRM($r = m - \ell, m - 1$) code,

$$d_{k-\gamma} \leq n - \gamma' \text{ where } k = \binom{m}{r}, \quad n = \sum_{i=r}^m \binom{m}{i}, \quad (4)$$

for all $\gamma \in [0, k)$. and γ' is as given in Theorem:5.4 for a given γ .

γ	ρ	\mathbb{P}	γ'	k	n
0	(0,0,0)	ϕ	0	10	26
1	(0,0,1)	$\{1,2\}$	1	9	25
2	(0,0,2)	$\{1,2\}, \{1,3\}$	2	8	24
3	(0,1,0)	$\{1,2,3\}$	4	7	22
4	(0,1,1)	$\{1,2,3\}, \{1,4\}$	5	6	21
5	(0,2,0)	$\{1,2,3\}, \{1,2,4\}$	7	5	19
6	(1,0,0)	$\{1,2,3,4\}$	11	4	15
7	(1,0,1)	$\{1,2,3,4\}, \{1,5\}$	12	3	14
8	(1,1,0)	$\{1,2,3,4\}, \{1,2,5\}$	14	2	12
9	(2,0,0)	$\{1,2,3,4\}, \{1,2,3,5\}$	18	1	8

TABLE I. Parameters list of $\text{SPRM}(2, 4, \gamma)$ code for $\gamma \in [0, 9]$. On counting the 2-element subsets of sets in \mathbb{P} gives γ and counting subsets of cardinality ≥ 2 gives γ' .

a) d_1 of *PRM* codes: For a $\text{PRM}(r = m - \ell, m - 1)$ code there are $\tau = 2^\ell$ disjoint recovery sets. This ensures that any $e \leq \tau - 1$ erasures can be corrected. This gives

$$d_1 = d_{\min} \geq 2^\ell.$$

Now by substituting $\gamma = k - 1$ in eq:4, we get an upper bound on d_1 . By the unique representation shown in Theorem:5.4, $\gamma = k - 1$ is represented by vector $(r, 0, \dots, 0)$. This gives:

$$\gamma' = h_1(r, \ell - 1) = \sum_{j=0}^{\ell-1} \sum_{i=0}^{r-1} \binom{r + \ell - 1 - i}{r + j - i} \quad (5)$$

It can be noted that

$$\sum_{i=0}^{r+j} \binom{r + \ell - 1 - i}{r + j - i} = \binom{r + \ell}{r + j} = \binom{m}{r + j}$$

Substituting the above equation in eq:5 we have

$$\begin{aligned} \gamma' &= \sum_{j=0}^{\ell-1} \binom{m}{r + j} - \sum_{j=0}^{\ell-1} \sum_{i=r}^{r+j} \binom{r + \ell - 1 - i}{r + j - i} \\ &= n - 1 - \sum_{j=0}^{\ell-1} \sum_{i=0}^j \binom{\ell - 1 - i}{j - i} = n - 2^\ell. \end{aligned}$$

This gives $d_1 = 2^\ell$.

b) d_2 of *PRM* codes: $\gamma = k - 2$ can be represented as $(r - 1, 1, 0, \dots, 0)$ when $m > r$ (i.e., $k > 1$). This gives

$$\begin{aligned} \gamma' &= h_1(r - 1, \ell - 1) + h_1(1, \ell - 2) \\ &= h_1(r, \ell - 1) - \sum_{j=0}^{\ell-1} \binom{\ell}{j + 1} + \sum_{j=0}^{\ell-2} \binom{\ell - 1}{j + 1} \\ &= n - 3(2)^{\ell-1} \end{aligned}$$

Substituting this in 4 we get $d_2 \leq 3(2)^{m-r-1}$.

VI. BOUNDS FOR SYSTEMATIC PIR CODES

For a systematic PIR code, the generator matrix is of the form $[I \mid P]$, where I is the $k \times k$ identity matrix. In this section we prove a lower bound on block length $n(k, \tau)$ of a systematic (n, k) τ -server PIR code. This is an improvement over the lower bound provided in [8]. We show in Section:VII that this bound is achieved for the case of $\tau = 3, 4$ by using $\text{PRM}(m-2, m-1)$ codes and their extensions.

Theorem 6.1: For a (n, k) 3-server systematic PIR code,

$$n(k, 3) \geq k + \left\lceil \frac{\sqrt{8k+1} + 1}{2} \right\rceil.$$

Proof: We consider a (n, k) 3-server systematic PIR code. For this code, let $R_{i1}, R_{i2}, \{i\}$ be the 3-disjoint recovery sets corresponding to message symbol i and let $\begin{bmatrix} I_k & g_{k+1} & \cdots & g_n \end{bmatrix}$ be the generator matrix G . Then,

$$e_i = \sum_{j \in S_{i1}} e_j + \sum_{j \in T_{i1}} g_j = \sum_{j \in S_{i2}} e_j + \sum_{j \in T_{i2}} g_j$$

where, $S_{i1} = R_{i1} \cap [k]$, $S_{i2} = R_{i2} \cap [k]$, $T_{i1} = R_{i1} \setminus S_{i1}$ and $T_{i2} = R_{i2} \setminus S_{i2}$. Let us define

$$\begin{aligned} u_{i1} &= \sum_{j \in S_{i1}} e_j, & u_{i2} &= \sum_{j \in S_{i2}} e_j, \\ v_{i1} &= \sum_{j \in T_{i1}} g_j = e_i + u_{i1}, & v_{i2} &= \sum_{j \in T_{i2}} g_j = e_i + u_{i2}. \end{aligned}$$

It is clear to see that,

$$e_i = (e_i + u_{i1}) \odot (e_i + u_{i2}) = v_{i1} \odot v_{i2} = \sum_{\substack{\ell \in T_{i1} \\ m \in T_{i2}}} g_\ell \odot g_m,$$

where \odot is the component wise product. Now consider set $X = \{g_{k+1}, \dots, g_n\}$ and define the set

$$X^2 = \{g_i \odot g_j \mid g_i, g_j \in X \text{ \& } i \neq j\}.$$

This gives $e_i \in \langle X^2 \rangle$ as $T_{i1} \cap T_{i2} = \emptyset$. Therefore we have,

$$k = \dim(e_1, \dots, e_k) = \dim(\langle X^2 \rangle) \leq |X^2| \leq \binom{n-k}{2}.$$

This gives us the bound for $\tau = 3$.

Corollary 6.2:

$$n(k, \tau) \geq k + \left\lceil \frac{\sqrt{8k+1} + 1}{2} \right\rceil + \tau - 3.$$

This corollary holds due to the fact that $n(k, \tau) \geq n(k, \tau-1) + 1$ since deletion of a column from the generator matrix will reduce τ by at most 1 (by [1]). Applying this fact to the bound $n(k, 3) \geq k + \left\lceil \frac{\sqrt{8k+1} + 1}{2} \right\rceil$ we get Corollary 6.2.

VII. OPTIMAL CODES FOR $\tau \leq 4$

For $\tau = 2$, $\text{PRM}(k-1, k-1)$ is the parity check code and it is optimal. To get a PIR code with dimension k and $\tau = 4$, consider $\text{PRM}(m-2, m-1)$ code, with m such that $k \in \left(\binom{m-1}{2}, \binom{m}{2}\right]$ and $\gamma = \binom{m}{2} - k$. By setting any γ message symbols to be zero, we can delete the coordinates corresponding to them. This gives:

$$k = \binom{m}{2} - \gamma, \quad n = k + m + 1, \quad \tau = 4.$$

In fact $\text{SPRM}(m-2, m-1, \gamma)$ has the same parameters as above. This gives $n(k, 4) \leq k + m + 1$. From the lower bound on block length in Corollary 6.2, we have that $n(k, 4) \geq k + m + 1$.

On puncturing $\text{SPRM}(m-2, m-1, \gamma)$ at a parity symbol we get a (n, k) 3-server PIR code, where $n = k + m$ and $k = \binom{m}{2} - \gamma$. This gives the upper bound $n(k, 3) \leq k + m$. From the lower bound in Theorem:6.1 we have $n(k, 3) \geq k + m$. Therefore for any k we have optimal systematic PIR codes for $\tau = 3, 4$.

In [2] it was shown that optimal systematic PIR codes for $\tau = 3, 4$ give optimal systematic primitive multi-set batch codes. So $\text{SPRM}(m-2, m-1, \gamma)$ and its punctured version can be used as $(n = k + m + 1, k = \binom{m}{2} - \gamma, 4)_B$, $(n = k + m, k = \binom{m}{2} - \gamma, 3)_B$ batch codes respectively.

k \ \tau	3*		4*		8		16	
	n_1	n_2	n_1	n_2	n_1	n_2	n_1	n_2
2	5	5	6	6	12	12	24	24
3	6	6	7	7	14	14	28	28
4	8	8	9	9	15	15	30	30
5	9	10	10	11	19	19	31	31
6	10	11	11	12	21	21	39	40
7	12	12	13	13	22	23	43	43
8	13	13	14	14	24	28	45	54
9	14	14	15	15	25	30	46	60
10	15	17	16	18	26	35	50	61
11	17	19	18	20	30	37	52	67
12	18	20	19	21	32	39	53	69
13	19	21	20	22	33	41	55	71
14	20	22	21	23	35	43	56	74
15	21	23	22	24	36	44	57	80
16	23	24	24	25	37	45	65	84
17	24	27	25	28	39	46	69	86
18	25	28	26	29	40	47	71	88
19	26	29	27	30	41	48	72	90
20	27	30	28	31	42	49	76	92
21	28	31	29	32	46	50	78	94
22	30	32	31	33	48	51	79	100
23	31	33	32	34	49	52	81	104
24	32	34	33	35	51	53	82	106
25	33	35	34	36	52	54	83	108
26	34	38	35	39	53	55	87	110
27	35	39	36	40	55	56	89	112
28	36	40	37	41	56	57	90	114
29	38	41	39	42	57	58	92	116
30	39	42	40	43	58	59	93	118
31	40	43	41	44	60	60	94	120
32	41	44	42	45	61	61	96	122

TABLE II. Block length for various k, τ . Here n_1 is the block length of the SPRM constructions and n_2 is the block length of the best known codes provided in [1]

REFERENCES

- [1] A. Fazeli, A. Vardy, and E. Yaakobi, "PIR with low storage overhead: Coding instead of replication," *CoRR*, vol. abs/1505.06241, 2015.
- [2] A. Vardy and E. Yaakobi, "Constructions of batch codes with near-optimal redundancy," in *IEEE International Symposium on Information Theory, ISIT*, 2016, pp. 1197–1201.
- [3] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *J. ACM*, vol. 45, no. 6, pp. 965–981, 1998.
- [4] N. B. Shah, K. V. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *IEEE International Symposium on Information Theory ISIT*, 2014, pp. 856–860.
- [5] D. Augot, F. Levy-dit-Vehel, and A. Shikfa, "A storage-efficient and robust private information retrieval scheme allowing few servers," in *Cryptology and Network Security - 13th International Conference, CANS*, 2014, pp. 222–239.
- [6] T. H. Chan, S. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *IEEE International Symposium on Information Theory, ISIT 2015, Hong Kong, China, June 14-19, 2015*, 2015, pp. 2842–2846.

- [7] A. Fazeli, A. Vardy, and E. Yaakobi, "Codes for distributed PIR with low storage overhead," in *IEEE International Symposium on Information Theory, ISIT*, 2015, pp. 2852–2856.
- [8] S. Rao and A. Vardy, "Lower bound on the redundancy of PIR codes," *CoRR*, vol. abs/1605.01869, 2016.
- [9] S. R. Blackburn and T. Etzion, "PIR array codes with optimal PIR rate," *CoRR*, vol. abs/1607.00235, 2016.
- [10] Y. Zhang, X. Wang, H. Wei, and G. Ge, "On private information retrieval array codes," *CoRR*, vol. abs/1609.09167, 2016.
- [11] D. E. Muller, "Application of boolean algebra to switching circuit design and to error detection," *Trans. I.R.E. Prof. Group on Electronic Computers*, vol. 3, no. 3, pp. 6–12, 1954.
- [12] I. S. Reed, "A class of multiple-error-correcting codes and the decoding scheme," *Trans. of the IRE Professional Group on Information Theory (TIT)*, vol. 4, pp. 38–49, 1954.
- [13] G. Lachaud, "Projective reed - muller codes," in *Coding Theory and Applications, 2nd International Colloquium*, 1986, pp. 125–129.